The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels

*Marylou Pausewang Gelfer and †Victoria A. Mikos

Milwaukee, Wisconsin, and Schaumburg, Illinois

Summary: The purpose of this study was to determine the accuracy with which listeners could identify the gender of a speaker from a synthesized isolated vowel based on the natural production of that speaker when (1) the fundamental frequency was consistent with the speaker's gender, (2) the fundamental frequency was inconsistent with the the speaker's gender, and (3) the speaker was transgendered. Ten male-to-female transgendered persons, 10 men and 10 women, served as subjects. Each speaker produced the vowels /i/, /u/, and /ɜ˞/. These vowels were analyzed for fundamental frequency and the first three formant frequencies and bandwidths. Formant frequency and bandwidth information was used to synthesize two vowel tokens for each speaker, one at a fundamental frequency of 120 Hz and one at 240 Hz. Listeners were asked to listen to these tokens and determine whether the original speaker was male or female. Listeners were not aware of the use of transgendered speakers. Results showed that, in all cases, gender identifications were based on fundamental frequency, even when fundamental frequency and formant frequency information was contradictory.

Key Words: Transgendered voice—Gender identification—Vowels—Formant frequencies.

INTRODUCTION

The most obvious difference between male and female voices is fundamental frequency or pitch. The average speaking fundamental frequency for men generally falls between 100 and 146 Hz, whereas the average speaking fundamental frequency for women is usually between 188 and 221 Hz.[1] These pitch levels help a listener correctly identify the speaker's gender. However, in addition to fundamental frequency, resonance might also play a role in gender identification.[2] Resonance is

Accepted for publication October 28, 2004.

From the *University of Wisconsin–Milwaukee, Milwaukee, Wisconsin; †Schaumburg School District #54, Schaumburg, Illinois.

Address correspondence and reprint requests to Marylou Pausewang Gelfer, Department of Communication Sciences and Disorders, University of Wisconsin–Milwaukee, P.O. Box 413, Milwaukee, WI 53201. E-mail: gelfer@uwm.edu

Journal of Voice, Vol. 19, No. 4, pp. 544–554
0892-1997/$30.00

doi:10.1016/j.jvoice.2004.10.006

a function of the supralaryngeal vocal tract. The air in the oral cavity, oropharynx, laryngopharynx, and (for some phonemes) the nasal cavities and nasopharynx vibrates at various frequencies in response to the vibratory movement of the vocal folds and air passing through the glottis. These resonant frequencies depend on the size and shape of the vocal tract and its constrictions (as well as tongue and lip positions, which can change the functional length of the vocal tract). Vocal tract resonances are often studied in terms of vowel formant frequencies. Because the male vocal tract is about 15% longer than the female vocal tract, the speech of men can be expected to have lower formant frequencies than those considered characteristic of women.[2]

The issue of vocal tract resonances, vowel formant frequencies, and the contribution of these elements to the perceived gender of a speaker is of special interest to transgendered persons. Those who perceive themselves as transgendered have a strong and persistent feeling that their biological gender is inappropriate for their psychological character. As a result of these conflicting traits, they might embark on a gender reassignment process to become recognized as a member of the gender they identify with.[3] Acquiring the voice characteristics of the reassigned gender is a particular challenge for male-to-female transgendered persons, because the vocal mechanism in most cases has attained adult male dimensions, and it is not affected by the administration of female hormones.[4]

Studies of the vocal characteristics of male-to-female transgendered persons who are successfully perceived as women have shown unequivocally that increasing speaking fundamental frequency is of primary importance when trying to shift perception of the voice from man to woman.[4–7] However, mixed results have been obtained regarding the importance of vowel formant frequencies in the vocal transition process. Mount and Salmon[7] studied the voice characteristics of a 63-year-old male-to-female transgendered person over 88 hours of speech therapy during an 11-month period. Therapy targeted an elevation of speaking fundamental frequency, use of a more anterior tongue carriage to raise the second formant of vowels (F2), and use of a more breathy quality. Mount and Salmon found that as fundamental frequency increased, in general, so did the second formant frequencies of the vowels of the client, although the rise in fundamental frequency preceded an elevation of formant frequencies. Moreover, they reported that the client was not perceived as a woman on the telephone until 10 months after the initiation of treatment, when elevated second formant frequencies were consistently used.

Gelfer and Schofield[5] investigated various acoustic measures of voice, including vowel formant frequencies, to determine which measures differentiated male-to-female transgendered persons still perceived as men from those perceived as their new gender, female. Their participants included 10 transgendered persons who were consistently perceived as men based on an audio-recorded speech sample, and 3 transgendered persons who were consistently perceived as women also based on an audio signal only. Results showed significant differences in speaking fundamental frequency and the upper limit of speaking fundamental frequency range between the perceived-male and perceived-female transgendered groups, with those who were perceived as women having the higher speaking frequencies and limits. However, no significant differences were found between the two transgendered groups for the first three formants of the vowels /i/ or /ɑ/, despite the observation that all formants were consistently higher in the perceived-female group.

Gelfer and Schofield did not confirm the impressions of Mount and Salmon regarding the importance of elevated formant frequencies in gender perception; however, their study had several limitations that leave open the question of the role of vocal tract resonances. First, their group of transgendered participants who were perceived as women included only 3 persons, compared with 10 in the perceived-male group. The number of perceived-female participants might have been too small to show significant results. Further, the transgendered persons who participated in the Gelfer and Schofield research were in varying stages of the transition process. Some had undergone speech therapy to develop a more feminine voice. Some were still living as men. Thus, even in the perceived-female group, the development of a feminine voice might have been incomplete.

An additional limitation common to all studies of formant frequency contribution toward gender

identification is the interaction of fundamental frequency and formant frequencies when natural vowels are used. In producing a fundamental frequency at a particular level, a person may use a certain vertical laryngeal height and tongue posture that affects vowel formants. The interaction between vocal fold vibratory patterns and supralaryngeal movements might be responsible for the gender cues listeners perceive. Manipulating one parameter while keeping the other constant to see how perception shifts is not possible when we use vowels produced naturally.

Some researchers have attempted to experimentally separate fundamental frequency from formant frequencies through the use of alternative voicing sources. For example, Schwartz and Rine[8] studied gender identification for the whispered vowels /i/ and /ɑ/. Results showed that listeners were extremely accurate in their gender identifications. No misidentifications of the gender of the speaker were made for /ɑ/, whereas four errors of gender identification were made for /i/. Spectral analysis showed that formants produced by women were higher in frequency than that by men. The findings of Schwartz and Rine supported the hypothesis that vowel formant frequencies can be cues for gender perception. However, it could be argued that a male whisper is acoustically different from a female whisper and that, in fact, voice source information as well as resonance information was available to listeners in identifying gender.

Coleman[9] eliminated the variable of voice source by having all of his speakers use an electrolarynx with a fundamental frequency of 85 Hz. In this investigation of gender identification, 10 male and 10 female speakers produced the vowels /i/ and /u/ in isolation and read portions of the Rainbow Passage using a Western Electric Company Model 5 electrolarynx (Lookout Mountain, TN). The readings of the Rainbow Passage were presented to 15 undergraduate speech students, who were asked to identify the speaker's gender and to rate their confidence in their selection. Results of the gender identification procedure showed that the gender of male speakers was correctly identified 98% of the time, whereas the gender of female speakers was correctly identified only 79% of the time. This difference in correct identifications might have been caused by the very low fundamental frequency of the electrolarynx, which was more appropriate to male speakers than to female speakers. Results of spectrographic analysis for formant frequencies in isolated vowels showed that, on the average, women had higher formant frequencies than men, but statistical analyses were not completed. It was concluded that, even with a single-frequency sound source (the electrolarynx) and a complete absence of vocal fold vibratory characteristics, correct gender identification was possible; therefore, some information regarding gender must be conveyed by formant frequencies or vocal tract resonances.

Coleman[10] further examined fundamental frequency and formant frequencies for their relative contributions to gender perception. Twenty men and 20 women served as speakers for the study. Each speaker produced portions of the Rainbow Passage and of four isolated vowels in three different conditions: (1) using his or her natural voice; (2) using an electrolarynx with a fundamental frequency of 120 Hz; and (3) using an electrolarynx with a fundamental frequency of 240 Hz. Twenty-five listeners then heard ten of the speakers, the five men and five women with the lowest (for men) or highest (for women) formant frequencies, reading 5 seconds of the Rainbow Passage in the two electrolarynx conditions. As in the first Coleman[9] study, listeners had to determine the gender of the speakers.

The Coleman study[10] was unique in that its design permitted a male-appropriate fundamental frequency (the 120-Hz electrolarynx) to be paired with female vowel formants and vocal tract resonances (produced by female speakers). Similarly, a female-appropriate fundamental frequency (the 240-Hz electrolarynx) could be paired with male vowel formants and vocal tract resonances (produced by male speaker subjects). Results showed that when the male fundamental frequency was paired with male vocal tract resonances, correct gender identification (male) was 100%. When the female fundamental frequency was paired with female vocal tract resonances, correct gender identification was not quite as high, but still a respectable 96%. When a mismatch between fundamental frequency and resonance characteristics occurred, the results seemed to favor men. When the male fundamental frequency was paired with female vocal tract resonances (that is, when

female speakers used the 120-Hz electrolarynx), the speaker was identified 67% of the time as a man; this indicates an important role for fundamental frequency in gender identification compared with resonance information. However, when a female fundamental frequency was paired with male vocal tract resonances (male speakers using the 240-Hz electrolarynx), the speaker was still identified 70% of the time as a man. Thus, Coleman[10] concluded that male cues, whether fundamental frequency or vowel formants, seemed to be more perceptually salient than females cues.

Although the Coleman[10] study provided interesting information on the perceptual results of a mismatch between fundamental frequency and formant frequencies, it also contained some methodological issues that could have influenced the results. First, speaker participants were selected on the basis of their vowel formant frequencies in isolated vowels, but listeners made their gender identifications based on the Rainbow Passage, in which the formants of speakers could have been very different because of coarticulatory factors. Thus, it is not certain whether the vowel formant frequencies of the two groups were actually different from one another. Second, because running speech was used to make the gender identifications, many acoustic cues other than vowel formant frequencies might have influenced listener judgments (for example, timing or intonation patterns). Finally, Coleman did not investigate the resonance characteristic of his electrolarynges. The spectrum of frequencies output by an electrolarynx is not flat, and the natural regions of energy minima and maxima of the electronic device could have interfered with listener perception of formant frequencies of the speakers. Thus, another type of voicing source would be beneficial.

Whiteside[11] used a matched/mismatched perceptual procedure similar to the Coleman[10] study but used synthesized vowels rather than an electrolarynx and the Rainbow Passage to determine the relative importance of fundamental frequency and formant frequencies in gender identification. Whiteside synthesized ten isolated vowels based on the natural productions in sentences. Her speakers were three men and three women. To derive "female" vowel formants, Whiteside averaged the formant frequencies produced for each vowel over her three female speakers. The "female" formants were then paired both with female-appropriate fundamental frequencies and male-appropriate fundamental frequencies. Formants determined from an average of those produced by male speakers were also paired with male-appropriate and female-appropriate fundamental frequencies. In contrast to the Coleman[10] listeners, who heard speakers read the Rainbow Passage with two different electrolarynges, Whiteside's listener subjects heard synthesized isolated vowels at two different fundamental frequencies.

The Whiteside results, like the Coleman results,[10] indicated that when male-appropriate formants were paired with a male-appropriate fundamental frequency, listeners almost always perceived gender correctly (97.2%, averaged over 10 vowels). When female-appropriate formants were paired with a female-appropriate fundamental frequency, gender identification was not as accurate (85.0%, averaged over 10 vowels); this is similar to Coleman's findings. When fundamental frequency-formant frequency mismatches occurred, the Whiteside results, unlike the Coleman[10] findings, showed the clear superiority of fundamental frequency as a cue for gender. A male-appropriate fundamental frequency paired with female-appropriate formants resulted in a gender identification of male 93.8% of the time. A female-appropriate fundamental frequency paired with male-appropriate formants resulted in gender identification of female 74.6% of the time. Clearly, fundamental frequency was the more salient cue for gender.

The Whiteside results emphasized the importance of fundamental frequency as a cue for speaker gender. However, because she synthesized her experimental stimuli from vowels produced naturally in running speech, vowel durations were extremely short (50 ms for short vowels such as /I/ and /æ/, and100 ms for long vowels such as /i/ and /u/). Further, her vowels were synthesized with a fundamental frequency contour, either rising or falling, depending on the intonation pattern of the sentence from which the vowel was excised. Some vowels of the women were consistently misjudged as "male" even when female formants were paired with female fundamental frequencies, and Whiteside hypothesized that falling fundamental frequency contours could have miscued listeners. Finally, the Whiteside subjects

spoke British General Northern English, and several vowels she studied are not typically produced in General American English.

A further question for research involves the contribution of fundamental frequency versus formant frequencies to the identification of transgendered subjects as their reassigned gender. Male-to-female transgendered speakers in particular might be hypothesized to have vowel formants appropriate for male speakers, given the size of their vocal tracts. However, in acquiring female voice characteristics, such persons may have learned articulatory and laryngeal postures to change vowel formant frequencies. Thus, gender identification based on isolated vowels (and the synthetic vowels created from them), which are produced by transgendered speakers with some success in developing a feminine voice, are also of interest.

This study was undertaken to replicate and expand the results of Whiteside[11] on a population of speakers of General American English, with isolated vowels of longer duration (250 ms) and a steady-state fundamental frequency to eliminate insufficient length and varying frequency contours as possible factors in listener judgment of gender. Specific research questions were are follows: (1) How accurately can listeners identify the speaker's gender from a synthesized isolated vowel created from a natural isolated vowel production, when the fundamental frequency is consistent with the speaker's gender (and presumably his or her formant frequencies)? (2) When a fundamental frequency appropriate for one gender is paired with formant frequencies produced by a speaker of the opposite gender, which cue is most salient in listener judgments of gender? (3) When the formant frequencies of male-to-female transgendered subjects using their "best feminine voice" are the basis for vowel synthesis, how is gender perceived as a function of fundamental frequency? The answers to these questions might help determine whether listeners rely predominantly on fundamental frequency for their judgments of gender or whether formant frequencies also play a role.

METHOD

Speaker subjects

Ten male-to-female transgendered persons were included in the speaker subject pool. To qualify for this study, a transgendered person had to meet the following criteria: (1) had to have begun hormone treatment, be living in the social role as a woman full time, or have completed sexual reassignment surgery before the initiation of the study; (2) had to report that she/he had been at least "somewhat successful" in developing a feminine voice, either through therapy or on her/his own; (3) had to produce a fundamental frequency during sustained vowel production of 165 Hz or higher; and (4) had to produce the vowels /i/, /u/, and /ɝ/ consistent with General American English in the opinion of the first author. The transgendered subjects were recruited from a local transsexual support group. Subjects selected for this study ranged from 23 to 57 years of age, with a mean of 43.2 years. Mean fundamental frequency for sustained vowel production was 216.7 Hz.

Twenty nontransgendered subjects, 10 biological men and 10 biological women, were also included. These subjects, matched to the transgendered subjects by age, were recruited from the University of Wisconsin-Milwaukee campus via posted advertisements or personal contacts. They were also screened to ensure that their productions of /i/, /u/, and /ɝ/ were consistent with General American English in the opinion of the first author. Male subjects ranged in age from 23 to 57 years, with a mean of 43.3 years; female subjects also ranged in age from 23 to 57 ears, with a mean of 43.2 years.

Speech and voice samples

Each speaker was seated in a quiet room with an ambient noise level of less than 60-dB SPL. Subjects were instructed to use comfortable conversational levels of pitch and loudness while sustaining the vowels /i/, /u/, and /ɝ/. These three vowels were selected based on the results of Peterson and Barney.[12] In their seminal study in the area of vowel acoustic structure and listener perception, Peterson and Barney noted that the vowels /i/, /u/, /æ/, and /ɝ/ showed the best intelligibility among listeners. They also found that /ɑ/ was often confused with /ə/ or /ɔ/, and it received low intelligibility scores. Thus, to permit the maximum intelligibility and discriminability among the synthesized samples, the vowels /i/, /u/, and /ɝ/ were ultimately selected for this study.

Transgendered speakers were instructed to produce the experimental stimuli using their best feminine voice. All subjects were asked to practice sustaining the three vowels for 5 seconds each. Once they were comfortable with the task (ie, producing the targets in a fluent and natural manner), the subjects were recorded by using a mouth-to-microphone distance of 2 in. Speech samples obtained from the male-to-female transgendered persons were collected at a transsexual support group meeting in southeastern Wisconsin. The speech samples were digitally recorded at the meeting by using a Toshiba Multimedia Notebook computer (Toshiba America, New York, NY) and the Speech Analysis Module of *Dr. Speech for Windows*, Version 3.0 (Tiger Electronics, Pawtucket, RI).[13] Speech samples obtained from the male and female subjects were recorded on a Gateway 2000 P5-75 computer, (Irvine, CA) also using the Speech Analysis Module of *Dr. Speech for Windows*, Version 3.0. For all subjects, a Radio Shack 33-30078 Electret unidimensional condenser microphone (Fort Worth, TX) was used to record speech samples. Samples were digitized at 44 kHz and stored for later analysis.

Vowel formant analysis

The three spoken vowels for each speaker were analyzed for the fundamental frequency and first three formant frequencies and bandwidths with the Speech Analysis Module of the *Dr. Speech* program. A fast Fourier transform (FFT) was conducted with a single 1024-point Hamming window. Linear predictive coding (LPC) analysis derived the frequencies and amplitudes of the lowest three formants by using 14 coefficients, with a 90-Hz preemphasis for male speakers and a 100-Hz preemphasis for female speakers. A 3-second segment from each vowel produced by each speaker was first analyzed by the second author. We used the most stable center portion of the vowel for analysis, excluding the onsets and offsets, to determine fundamental and formant frequencies and bandwidths. The first author independently analyzed each vowel for reliability. Analysis of the results of the two investigators showed agreement on 90% of the formants. The remaining discrepancies were resolved through joint reanalysis of the vowels in question.

Analysis of the vowel formant data showed that the mean of each formant frequency for each vowel for the female speakers was higher than the corresponding mean of each formant frequency for each vowel for the male speakers (Table 1). The formants derived in this study were similar to the values of both Hillenbrand et al[14] and Whiteside[11] (except for /ɝ/, which is produced differently by speakers of British English); this suggests that the present procedure for vowel formant analysis yielded valid results. The mean of the formant frequencies for each vowel for the transgendered speakers was sometimes higher and sometimes lower than its cognate from the male productions. For example, all three formants of /u/ were higher in the productions of transgendered speakers than they were for male speakers, as well as F1 and F2 of /i/. On the other hand, the first formant of /i/ and all three formants of /ɝ/ were lower than the corresponding average formants for men. Differences were apparent between the formant frequencies of men and women; and these differences conformed to expectations based on previous literature. The relationship between formants produced by biological men and male-to-female transgendered persons was less consistent.

Vowel synthesis and tape construction

Two synthesized vowel files were created for each spoken vowel for each subject with the Voice Synthesis and Therapy Module of the *Dr. Speech* program. We used a digitization rate of 44.1 kHz; thus, frequencies of 0 to 22.05 kHz were available during the synthesis process. The vowels were synthesized with the three formant frequencies and their corresponding bandwidths obtained during the analyses of vowels of each speaker. The glottal source for the vowel synthesis had a preselected amplitude of 60 dB and used a linear flat contour pattern over the 250-ms duration of each vowel. Rise and fall time for each vowel was 40 ms. One file was synthesized with a fundamental frequency of 120 Hz, whereas a second file with identical formants and bandwidths was synthesized with a fundamental frequency of 240 Hz. This process resulted in a total of 180 synthetic vowel tokens (30 speakers x 3 vowels x 2 fundamental frequencies).

TABLE 1. *Average Vowel Formants (in Hertz) for Male Subjects (N = 10), Female Subjects (N = 10), and Transgendered Subjects (N = 10)*

	Male	Transgendered	Female
/i/ F1	283.16	272.40	323.01
F2	2200.71	2365.44	2614.15
F3	2770.79	3138.77	3230.26
/u/ F1	295.00	305.23	372.53
F2	827.95	916.27	986.86
F3	2307.29	2421.40	2694.88
/ɜ˞/ F1	437.12	327.31	473.71
F2	1271.55	1066.99	1482.57
F3	2374.03	2146.54	2426.97

The synthesized vowels were recorded in random order on a stimulus tape in nine blocks of 20 vowels each. Each synthesized vowel was preceded by an identification number and followed by a 3-second response time. Listeners were first given a training block to familiarize them with the quality of the vowels. Reliability was assessed by randomly choosing three blocks (33% of the experimental stimuli) and including them a second time on the tape. Thirteen blocks (one training, nine experimental, and three reliability) were presented. Therefore, each listener heard a total of 260 synthetic vowels. The synthetic vowels were clearly recognizable as synthetic and not natural.

Listener subjects

The stimulus tape was presented to 30 normal-hearing young adult listeners (15 men and 15 women) ranging from 18 to 35 years of age, with a mean of 24.4 years. Listeners were recruited from undergraduate allied health professions, psychology and education classes, or through personal contacts. No subject had more than two courses in communication sciences and disorders, and none had performed coursework pertaining to speech science or voice disorders. All listeners passed a hearing screening in a sound-treated Industrial Acoustics Corporation (IAC) (Staines, Middlesex, United Kingdom) booth at 25 dB (HL) at 1 kHz, 2 kHz, and 4 kHz.

Perceptual protocols

Listeners were seated in small groups of one to three persons in an IAC booth. Before listening to the stimulus tape, the listeners were instructed via audiotaped instructions that the speech productions they would hear had been produced by a computer, and that the purpose of the study was to determine the acoustic cues listeners use to identify a speaker's gender. In addition, they were instructed to identify each speaker as a man or a woman, and rate how confident they were in their judgment based on a five-point scale (1 = guessing, 5 = very confident). All listeners were unaware of the use of transgendered speakers. The listening task took approximately 1 hour.

Listener reliability

As described above, three blocks of stimuli were presented to the listeners twice on the stimulus tape to assess intrajudge reliability. To remain in the study, all listener subjects had to obtain a reliability score of at least 50% (well above the chance level of matching gender identifications, calculated to be 25%). The overall average reliability for the group was 77.7%. That is, the speaker's gender that was identified on the first presentation of a vowel was the same that was identified by the listener on the second presentation of the vowel an average of 77.7% of the time.

Validity of the synthesized vowels

To assess whether the synthesized vowels were valid representatives of the original tokens they were modeled on, 60 of the 180 synthesized vowels were randomly selected for input into the Speech Analysis Module of the *Dr. Speech* program and subjected to FFT and LPC analysis. We used the same settings that had previously been used for the natural speech sample analyses. The output provided an F1, F2, and F3 measure for each synthesized vowel. These formants were then correlated with the F1, F2, or F3 obtained for the corresponding natural vowel.

Results of the correlations can be seen in Table 2. In general, the formants calculated for the synthetic vowels matched well with the formants measured in their natural syllable counterparts, with correlation coefficients ranging from $r = 0.797$ to $r = 0.999$, and significance levels ranging from $P = .006$ to $P = .000$. The two poorest correlation coefficients were between the first formant of the vowels of transgendered speakers and the first formant of

TABLE 2. *Results of Correlational Analyses Between Formant Frequencies Calculated from Natural Syllables and Formant Frequencies Calculated from Synthetic Syllables*

First formant (F1) from:
Transgendered samples X synthesized samples (SFF of 240 Hz): $r = .821$, $P = .004$
Transgendered samples X synthesized samples (SFF of 120 Hz): $r = .946$, $P = .000$
Male samples X synthesized samples (SFF of 240 Hz): $r = .948$, $P = .000$
Male samples X synthesized samples (SFF of 120 Hz): $r = .924$, $P = .000$
Female samples X synthesized samples (SFF of 240 Hz): $r = .797$, $P = .006$
Female samples X synthesized samples (SFF of 120 Hz): $r = .954$, $P = .000$
Second formant (F2)
Transgendered samples X synthesized samples (SFF of 240 Hz): $r = .987$, $P = .000$
Transgendered samples X synthesized samples (SFF of 120 Hz): $r = .994$, $P = .000$
Male samples X synthesized samples (SFF of 240 Hz): $r = .992$, $P = .000$
Male samples X synthesized samples (SFF of 120 Hz): $r = .999$, $P = .000$
Female samples X synthesized samples (SFF of 240 Hz): $r = .997$, $P = .000$
Female samples X synthesized samples (SFF of 120 Hz): $r = .995$, $P = .000$
Third formant (F3)
Transgendered samples X synthesized samples (SFF of 240 Hz): $r = .985$, $P = .000$
Transgendered samples X synthesized samples (SFF of 120 Hz): $r = .982$, $P = .000$
Male samples X synthesized samples (SFF of 240 Hz): $r = .970$, $P = .000$
Male samples X synthesized samples (SFF of 120 Hz): $r = .974$, $P = .000$
Female samples X synthesized samples (SFF of 240 Hz): $r = .985$, $P = .000$
Female samples X synthesized samples (SFF of 120 Hz): $r = .969$, $P = .000$

Note: For each correlation, we used 10 randomly selected natural vowels and 10 synthetic vowels based on those natural vowels.

the synthesized version of those vowels when a fundamental frequency of 240 Hz was used, and between the first formant of vowels of the female speakers vowels and the first formant of the synthesized version of those vowels when a fundamental frequency of 240 Hz was used. In both cases, it is probable that the high fundamental frequency affected the ability of the program to accurately identify the first formant (especially for /i/ and /u/) in the analysis of both the natural and the synthetic vowels.

The results of the correlational analyses between the formants of natural vowels of the speakers and their synthetic counterparts lend support to the validity of the synthesized stimuli. The synthetic stimuli were intended to represent natural stimuli, and it seems that that aim was achieved, at least in terms of vowel formant frequencies.

Statistics

A 3 × 2 analysis of variance (ANOVA)[15] was planned to assess listeners' percentage of correct gender identifications based on vowel formant frequencies (as produced by male, female, and transgendered subjects intended to be perceived as women) and fundamental frequency (120 Hz, 240 Hz). If the formant frequency main effect or the formant frequency × fundamental frequency interaction was found to be significant, Scheffé post hoc comparisons were planned. A .05 level of significance was selected.

Before statistical analysis, an independent perceptual analysis of each vowel stimulus by both investigators was undertaken. If each vowel was not recognizable as the intended phoneme to both investigators, that vowel was dropped from the analysis; this eliminated vowel intelligibility as a confounding factor in the gender identifications. The final data set retained for statistical analysis included 29 synthetic vowels with a fundamental frequency of 120 Hz (of a total of 30) and 28 synthetic vowels with a fundamental frequency of 240 Hz based on the productions of the male speakers, 21 synthetic vowels at 120 Hz and 24 synthetic vowels at 240 Hz based on the productions of the female speakers, and 20 synthetic vowels at 120 Hz and 14 synthetic vowels at 240 Hz based on the productions of the transgendered speakers. The remaining vowels and associated listener responses were eliminated from

the data set because of difficulty with vowel intelligibility.

RESULTS

Identification results

Of the 29 synthetic vowels combining male-appropriate formant frequencies and a male-appropriate fundamental frequency, 84.2% were accurately judged to be produced by male speakers (Table 3). Of the 28 synthetic vowels combining male-appropriate formant frequencies with a female-appropriate fundamental, only 19.3% were correctly judged to be produced by men.

For female speakers, of the 24 synthetic vowels with female-appropriate formant frequencies and a female-appropriate fundamental frequency, 73.8% were correctly identified as being produced by a female speaker. Of the 21 synthetic vowels combining female-appropriate formants with a male-appropriate fundamental, only 20.2% were identified as being produced by a woman.

The identification results for transgendered speakers were similar to the data for female speakers. When the vowel formants of transgendered speakers using their best feminine voice were paired with a female-appropriate fundamental frequency, 84.0% of the 14 included vowels were judged as being produced by a woman. When the same female-intended vowel formants were paired with a male-appropriate fundamental frequency, only 16.7% of the 20 included vowels were identified as being produced by a woman.

Comparisons between groups

An ANOVA was used to determine whether there were significant differences among percentages of correct means based on vowel formant frequencies (male, female, and male-to-female transgendered), fundamental frequency (120 Hz, 240 Hz), or an interaction between the two. Analysis showed that there was a significant interaction ($P < .05$) between fundamental frequency and formant frequencies (Table 4). That is, when formant frequencies appropriate to the actual or intended gender of the speaker were paired with pitch appropriate to gender (or intended gender), listener subjects identified gender with significantly higher accuracy than when formant frequencies and pitch were mismatched.

TABLE 3. *Percent Correct Gender Identifications at the Two Different Fundamental Frequencies for the Vowel Stimuli Presented to Listeners*

	Speaking Fundamental Frequency	
	120 Hz	**240 Hz**
Group		
Biological Men	84.2% (N = 29)	19.3% (N = 28)
Biological Women	20.2% (N = 21)	73.8% (N = 24)
Transgendered Participants	16.7% (N = 20)	84.0% (N = 14)

Note: For each cell, total number of vowel tokens included in the analysis is reported (N).

Post hoc Scheffé tests[16] were used to determine whether there were any other significant comparisons based on vowel formant frequencies, particularly between female and transgendered speakers. When female and transgendered speakers using a female-appropriate fundamental frequency were compared, there were no significant differences at the .05 level in identification accuracy of the samples as being produced by a woman. Similarly, when female and transgendered speakers using a male-appropriate fundamental frequency were compared, again there was no significant difference in identification accuracy. Both were inaccurately identified most of the time. Finally, because it seemed that men might have been identified with greater accuracy than women, a male formant-male fundamental frequency versus female formant-female fundamental frequency comparison was made. This result too proved to be nonsignificant at the .05 level.

TABLE 4. *A 3 × 2 Analysis of Variance for Percent Correct Gender Identification as a Function of Vowel Formants (Men, Women, Transgendered Persons) and Fundamental Frequency (120 Hz, 240 Hz)*

Source	Sum of Squares	*df*	Mean Square	F	*P*
Vowel formants	725.349	2	362.674	1.171	.312
Fundamental frequency	15683.734	1	15683.734	50.641	.000
Vowel formants × Fundamental frequency	158619.361	2	79309.681	256.083	.000

DISCUSSION

One purpose of this study was to determine how accurately listeners judge gender based on synthetic vowels when the fundamental frequency of the vowel is consistent with the speaker's gender and vowel formant frequencies. Results showed that, for both male and female speakers, identification accuracy was relatively high when the two sets of cues were consistent (84.2% and 73.8%, respectively). Women were identified less accurately than men, but the difference was not significant.

A second purpose of this study was to determine the relative strength of fundamental frequency cues compared with formant frequency cues when the two were mismatched. Results of this study showed unequivocally that fundamental frequency cues were more salient to listeners than were formant frequency cues. For example, when the formants of a man were paired with a female-appropriate fundamental frequency, listeners perceived a male speaker only 19.3% of the time. When the formants of a woman were paired with a male-appropriate fundamental frequency, listeners perceived a female speaker only 20.2% of the time. These results suggest that formant frequency cues do not contribute strongly to gender identification, at least in isolated vowels.

A third purpose of this study was to explore the use of fundamental frequency versus formant frequency cues in gender identification of male-to-female transgendered speakers using their best feminine voices. Results showed that, for transgendered speakers as well as for biological men and women, fundamental frequency was the most salient cue to gender identification. Transgendered speakers' formant frequencies combined with a female-appropriate fundamental frequency resulted in a higher percentage of "female" identifications (84.0%) than was seen for biological women when their formant frequencies were combined with a female-appropriate fundamental (73.8%), although the difference did not reach statistical significance. It was hypothesized that this result occurred because the transgendered speakers were offered time to "warm up" so that they could use their best feminine voices and articulation patterns when repeating the target vowels. Biological female speakers, by contrast, might have used less-precise articulatory movements because no vocal changes or special practice was necessary.

The results of this study agreed in many respects with the results of Whiteside[11] and Coleman.[10] All three studies found that the highest percentage of correct listener responses was obtained when male formant frequencies were paired with a male-appropriate fundamental frequency. In all three studies, a lower percentage of correct gender identifications was made by listeners when female formant frequencies were paired with a female-appropriate fundamental frequency. Finally, both the Whiteside study and this study demonstrated that speaking fundamental frequency was clearly the dominant cue when fundamental frequency and formant frequencies were mismatched. These similarities are remarkable considering the different stimuli and methods used in the different studies: male and female speakers reading portions of the Rainbow Passage using two different electrolarynges of 120 and 240 Hz each in the Coleman[10] study; ten short-duration vowels (50 ms or 100 ms) synthesized with the averaged formants of three male and three female speakers with male-appropriate and female-appropriate fundamental frequency contours in the Whiteside study; and synthesized vowels of 250 ms duration based on three vowel productions of 10 male, 10 female, and 10 transgendered subjects with steady-state fundamental frequencies of 120 Hz and 240 Hz in this study.

There were differences among studies as well. The highest percentages of correct gender identifications were obtained by Coleman,[10] which might be expected because listeners heard connected speech samples. The Coleman listeners received not only vowel formant cues from the speaker subjects, but spectral cues from consonant production, intonation contours, timing and duration cues, and undoubtedly other cues as well. Whiteside[11] obtained correct gender identifications somewhat lower than those of Coleman, but the Whiteside subjects heard only isolated vowels. Some coarticulatory cues might have been present, because these isolated vowels were extracted from running speech, but much of the acoustic information available to the Coleman subjects was absent in the Whiteside stimuli. In fact, the fundamental frequency contours included in the stimulus set might have enhanced gender perception

for some vowels, although, according to Whiteside, it confounded it in other cases.

This study found correct gender identifications somewhat lower than those of Whiteside, but that may be due to the nature of the stimuli. Whiteside synthesized vowel formants were based on a composite of all speaker subjects, whereas we used the vowel formants of each speaker as the basis for a pair of synthetic vowel stimuli (one with a fundamental frequency at 120 Hz, one at 240 Hz). Some speakers' formants were easy to analyze and incorporate into a synthesized vowel, whereas others were more difficult to derive, and thus the corresponding synthetic vowel may have sounded less natural. Unintelligible vowels were excluded from the data set, but variation in vowel quality might have contributed to greater difficulty with gender identification.

Unfortunately, this study does not resolve the issue of the importance of formant frequencies and, more generally, vocal tract resonances, to male-to-female transgendered speakers attempting to be vocally recognized as a member of the reassigned gender. The transgendered speakers in this study, all of whom were (reportedly) successful at least some of the time in being vocally recognized as a woman, had higher average vowel formants than did the biological men for 5 out of 9 formants in this study (3 vowels $\times$ 3 formants each). In the process of adopting more feminine speech patterns, there seemed to be some shift in vocal tract resonance characteristics among the transgendered speakers. How necessary that shift is to changing overall gender perception remains unknown.

It is clear that, for isolated vowels, fundamental frequency is the most important cue to the gender perception of a speaker. However, other cues, both segmental (eg, consonant spectra) and supersegmental (eg, intonation patterns) might also play a role, based on the results of Coleman.[10] Further research at the word and sentence level is needed to determine the relative importance, if any, of these discrimination parameters, so that speech-language pathologists can better serve persons who need to enhance or alter the perceived gender of their voices.

REFERENCES

1. Baken RJ, Orlikoff RF. *Clinical Measurement of Speech and Voice*. 2nd ed. San Diego, CA: Singular Publishing Group Thomson Learning; 2000.
2. Bachorowski JA, Owren MJ. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *J Acoust Soc Am*. 1999;106(2):1054–1063.
3. Brown ML, Rounsley CA. *True Selves: Understanding Transsexualism*. San Francisco, CA: Jossey Bass Publishers; 1996.
4. Wolfe VI, Ratusnik DL, Smith FH, Northrop G. Intonation and fundamental frequency in male-to-female transsexuals. *J Speech Hearing Dis*. 1990;55:43–50.
5. Gelfer MP, Schofield KJ. Comparisons of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female vs. those perceived as male. *J Voice*. 2000;14:22–33.
6. Spencer LE. Speech characteristics of male-to-female transsexuals: a perceptual and acoustic study. *Folia Phoniat*. 1988;40:31–42.
7. Mount KH, Salmon SJ. Changing the vocal characteristics of a postoperative transsexual patient: a longitudinal study. *J Communication Dis*. 1988;21:229–238.
8. Schwartz MF, Rine HE. Identification of speaker sex from isolated whispered vowels. *J Acoust Soc Am*. 1968;44(6): 1178–1179.
9. Coleman RO. Male and female voice quality and its relationship to vowel formant frequencies. *J Speech Hearing Res*. 1971;14:565–577.
10. Coleman RO. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *J Speech Hearing Res*. 1976;19:168–180.
11. Whiteside SP. The identification of a speaker's sex from synthesized vowels. *Percept and Mot Skills*. 1998;87: 595–600.
12. Peterson GE, Barney HL. Control methods used in a study of the vowels. *J Acoust Soc Am*. 1952;24:175–184.
13. Huang D, Lin S, O'Brian R. *Dr. Speech for Windows*, Version 3.0 [Computer software]. Seattle, WA: Tiger Electronics, Inc.; 1995.
14. Hillenbrand J, Getty L, Clark MJ, Wheeler K. Acoustic characteristics of American English vowels. *J Acoust Soc Am*. 1995;97(5):3099–3111.
15. *Systat 8.0 for Windows* [Computer software]. Chicago, IL: SPSS Inc.; 1998.
16. Shearer WM. *Research Procedures in Speech, Language, and Hearing*. Baltimore, MD: Williams and Wilkins; 1982: 142–143.